Abstract Title Page
Not included in page count.

Title: Rapid and iterative estimation of predictions of high school graduation and other milestones

Authors and Affiliations: Kristin E. Porter, MDRC; Rekha Balu, MDRC; Brad Gunton, NVPS; Jefferson Pestronk, NVPS; Allison Cohen, NVPS

Abstract Body

Limit 4 pages single-spaced.

Background / Context:
Description of prior research and its intellectual context.

Recent work by Robert Balfanz, the Consortium on Chicago School Research and other education researchers have proposed Early Warning Systems and other indicator-based efforts to identify students at risk of not graduating (Balfanz, 2008; Bruce et. Al., 2011; Fairchild et al., 2011; Roderick et al., 2014; Mourshed et al., 2010). An indicator-based approach is suitable when schools or districts have data from limited points in time that provide a snapshot of student behavior and course performance, need dichotomous variables to classify students, or when they believe that only readily available variables are critical to monitor.

However, with the advent of data systems that allow for frequent or even real-time student data updates, and recognition that high school students often can move from being on-track to graduation to off-track in a matter of weeks, indicator analysis alone may not provide a complete picture to guide school leaders' actions. We suggest that approaches that capitalize on high-frequency data updates and treat risk as a more continuous measure can add more value. Moreover, models that also account for more nuanced and typically unobserved factors related to student risk may better identify students who need intervention. Predictive modeling methods, often used in business and medical research, can capitalize on these new, richer datasets and use data-adaptive methods to maximize flexibility. Yet very little has been written on ways to apply such methods to student-level education data to fill gaps left by simpler parametric approaches, and how they might inform practice decisions in timely and useful ways.

To this end, our organizations, a nonprofit organization that supports district-run public high schools in New York City and an independent research organization, are collaborating on a variety of analyses aimed at providing real-time evidence to school administrators, teachers and our support organization partner about which students are most at risk of not graduating and of not meeting key milestones on the path to graduation. We have been using a large-scale student-level dataset from 77 New York City high schools with weekly and sometimes daily updates on student attendance and performance on courses and gateway exams, creating an important testing ground for making use of multiple measures and for these modeling approaches. The nascent literature on research-practice partnerships (Tseng, 2012) emphasizes the importance of co-creation and joint learning. In our case, both organizations jointly identified research questions and are learning together whether predictive approaches used in other fields add value, and what are critical decisions involved in the process of turning research into action.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

This paper will focus on how the addition of an iterative predictive modeling framework that can be used with real-time data updates might add value over simpler and more static methods and fill a gap in understanding when, how and for which students the risk of not graduating changes. This paper will be not an exhaustive review or a technical description of machine learning algorithms, because this is covered well in other statistical literature. Establishing an iterative framework is essential for several reasons: First, it creates a repeatable process that allows us to update these models as contextual factors change or new information becomes available; our data system is constantly updated with current values and new measures

as more data sources are incorporated (for example, period attendance or college readiness measures will soon be added). Additionally, the predictive modeling framework can be applied to multiple research questions, as is our plan.

Our goal is to use rapid and iterative data analyses that tell school leaders, for each student at a point in time, the likelihood of graduation and of meeting milestones required for graduation (advancing to the next grade, passing a course, passing an exit exam, etc.).We aim to answer the following research questions: For an individual student, what is the predicted likelihood that he or she will graduate on time? What is the uncertainty around this predicted likelihood? How does the predicted likelihood change over time (e.g., from one grade-level to the next as well as during each grade level)? How do predicted likelihoods of on-time graduation vary - within schools, across schools and across subgroups of students? How do predicted likelihoods of on-time graduation compare to actual graduation outcomes? What may explain false positives and false negatives (i.e. predictions of graduating on time or not graduating on time that prove to be incorrect)? We then asked similar questions for key milestones on the path to graduation. For example, what is the predicted likelihood that a student will pass a particular Regents exam?

Setting:
Description of the research location.

Our support organization partner has been providing support to public schools in New York City since 1989. The organization believes that focusing on the key systems involved in operating a school and applying a rigorous, data-driven mindset can manage the inherent complexity of high school and lead to greater high school and post-secondary success for our students.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.

Our support organization partner serves over 45,000 students in its 77 schools, which have a graduation rate of about 78 percent. One in five students is in a special education program, slightly higher than the city average, and students come in to high school with a higher level of need than in a typical New York City high school. Nearly three out of four students are either black or Hispanic, and about three quarters are eligible for free or reduced lunch as well.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

For New York City high schools students, passing five exit exams spanning English, Math, Science and Social Studies is a necessary pre-condition for graduation. Because many of our students enter high school below grade level, passing these exams is a challenge for many. While students may retake exams, the first failure has a ripple effect in terms of future course programming that may start to diminish the likelihood of on-time graduation. Knowing in advance which students are likely to fail an exam before their first attempt would allow us to focus resources and support to the schools and students who need it most. For example, identifying students at risk of failing their first attempt could facilitate communication with families regarding the importance of preparation, invitations to after school tutoring sessions, and

targeted mock exams. Preventing students from falling off track in the first place will be the most successful pathway towards on time graduation and post-secondary success.

Research Design:
Description of the research design.

We are not testing an intervention, but applying a new modeling approach.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

The support organization in our partnership has assembled a master dataset from multiple DOE sources that provides a rich (and constantly growing) set of measures that can be valuable in predicting key outcomes for students. As mentioned above, our goal was to develop a framework for answering multiple prediction questions rapidly and iteratively. Below we summarize our framework, which consists of nine analytic steps.

I. Define outcome and population of interest.

This defines the prediction problem of focus. Currently, our outcome of interest is on-time graduation and our population of interest is 4^{th} -year high school students..

II. Define subpopulations for whom there is particular interest

This defines a particular group for which there is particular focus for policy/practice reasons. Our current is particularly focused on students who could potentially graduate but who are at risk of not graduating because they are near the threshold of meeting graduation requirements.

III. Identify samples of the populations and subpopulations for building a predictive model.

Next, we need to decide on samples (for the populations defined in I and II) for which outcome data are available.

IV. Specify candidate predictors

This is the most important step and takes the most thought and work. Drawing on thousands of raw measures, we (1) *identify and create measures that we expect are predictive of the outcome of interest* based on substantive knowledge, even if only predictive for some subgroup of students (i.e. there may be an interaction effect); (2) s*elect or define measures in such a way that we expect their meaning to remain the same over time* in order to maximize generalizability of our model to future cohorts; (3) *include school indicators* because they can be valuable in explaining variation due unobservable characteristics; (4) *consider composite measures* that would allow us to add more information to the modeling without overwhelming it with too many variables (since we have thousands); (5) *define the measures in a way that includes missing values* so that we don't exclude any students from the modeling.

V. Specify candidate model "learners."

A "learner" refers to either a pre-specified parametric regression model or machine learning algorithm that we can "train" (i.e. develop and fit) so that it results in a final predictive model. We recommend starting with a simple parametric model in which a specified set of predictors are included as single terms. If the set of candidate predictors is a good one, this simple model may work very well – resulting in a good fit and highly accurate predictions. However, there may be missed opportunities for extracting more information from the data – either in terms of loosening the linearity assumption, of allowing for interaction effects, and of increasing generalizability. Therefore, we have been implementing a variety of machine learning algorithms for a binary outcome, including penalized logistic regression, stepwise logistic regression, the k-nearest

neighbor (k-nn) algorithm, the random forest algorithm, and the Chi-Square Automatic Interaction Detector (CHAID) algorithm.

VI. Compare the performance of the candidate learners and pick the best one

We want to evaluate each of the learners and compare across them in terms of two concepts: (1) "calibration", which refers to how well a learner fits the data, and (2) "discrimination," which refers to the accuracy of the learner in ranking of predicted probabilities correctly and thereby classifying individuals into categories correctly. We consider multiple measures of these two concepts (e.g. goodness of fit, sensitivity, specificity). We compute each measure in such a way that we are evaluating how well the learners would perform in *new samples* from the same population. We want the learners to be generalizable to new samples, as we will use them to predict outcomes in future cohorts of students. Therefore, we use *v*-fold cross-validation.

VII. Using a new sample (for which the outcome of interest may or may not be known), compute estimated predicted probabilities and estimate uncertainty.

Using the selected "best" learner, based on some combination of the cross-validated performance measures, we next train/fit that learner on the entire dataset. The resulting estimated model is then applied to a new sample to obtain predicted probabilities for each individual in the new sample. To estimate this uncertainty, we have implemented a nonparametric bootstrap procedure.

VIII. Interpret findings

Our predictive modeling produces, for each student in a given cohort, a predicted likelihood of on-time graduation and an interval around that prediction that estimates the uncertainty due to sampling variability. Because the predicted probability depends on a model fit on previous cohorts, a limitation and caution in interpreting the predicted probabilities is that there is additional and unquantifiable uncertainty due to *changes* in factors that would predict students' likelihood of on time graduation. It may be most useful to rank students by their predicted probabilities in order to identify students who could be prioritized for a next step (e.g., for further investigation about their situation or an intervention). In addition, we can examine distributions of predicted probabilities within and across schools and for key subgroups. We can also examine trends over time.

IX. Iterate - with same population and outcome of interest – as new information becomes available

Our models are dynamic rather than static. New or updated measures that may be good predictors of our outcome of interest will become available after we have completed this entire process. We plan to add new predictors and update our model selection, fit and predictions. Intervening on students in the analyses could (and should) also affect relationships in the model and thus prompt iteration (with or without new measures).

Findings / Results:

Description of the main findings with specific details.

We can disclose findings once they are more formally shared with schools and the NYC DOE.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

We will describe benefits and limitations of the iterative framework, compare it to indicator frameworks, and make recommendations for practice. We will discuss next steps – other prediction problems and follow-up analyses that help us better understand the prediction results, as well as interventions that our analyses suggest.

Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.

Balfanz, Robert (2008). Three Steps to Building an Early Warning and Intervention System for Potential Dropouts. Johns Hopkins University.

Bruce, Mary, John M Bridgeland, Joanna Hornig Fox, and Robert Balfanz (2011). On Track for Success: The Use of Early Warning Indicator and Intervention Systems to Build a Grad Nation. Washington, DC: Civic Enterprises.

Fairchild, Susan, Brad Gunton, Beverly Donohue, Carolyn Berry, Ruth Genn, and Jessica Knevals (2011). Student Progress to Graduation in New York City High Schools. New York, NY.

Mourshed, Mona, Chinezi Chijioke, and Michael Barber. 2010. How The World's Most Improved School Systems Keep Getting Better. London.

Roderick, Melissa, Thomas Kelley-Kemple, David W. Johnson, and Nicole O. Beechum. 2014. Preventable Failure: Improvements in Long-Term Outcomes when High Schools Focused on the Ninth Grade Year. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Tseng, Vivian (2012). The Uses of Research in Policy and Practice. Social Policy Report. 26:2.

Appendix B. Tables and Figures
Not included in page count.